УДК 004.67

Деревлев Никита Андреевич

студент, Белгородский государственный национальный исследовательский университет,

Россия, г. Белгород

Паманин Максим Юрьевич

студент, Белгородский государственный национальный исследовательский университет,

Россия, г. Белгород

Путивцева Наталья Павловна

кандидат технических наук, Белгородский государственный национальный исследовательский университет,

Россия, г. Белгород

БОЛЬШИЕ ДАННЫЕ И ИХ ПРИМЕНЕНИЕ В ЭКОНОМЕТРИКЕ

Аннотация: Настоящая научная статья посвящена исследованию применения больших данных в области эконометрики. В современной экономике объем и разнообразие доступных данных значительно возрастает, и требуется разработка новых методов анализа, которые позволят эффективно использовать эту информацию для выявления закономерностей и принятия более точных и обоснованных экономических решений.

Ключевые слова: Большие данные, Эконометрика, Информационные технологии, Машинное обучение, Прогнозирование, Анализ финансовых рынков, Бизнес-аналитика.

Derevlev Nikita Andreevich

student, Belgorod State National Research University,

Russia, Belgorod

Pamanin Maxim Yurievich

student, Belgorod State National Research University,

Russia, Belgorod

Putivtseva Natalia Pavlovna

Candidate of Technical Sciences, Belgorod State National Research University,

Russia, Belgorod

BIG DATA AND ITS APPLICATION IN ECONOMETRICS

Abstract: This scientific article is devoted to the study of the application of big data in the field of econometrics. In the modern economy, the volume and variety of available data is increasing significantly, and it requires the development of new analysis methods that will effectively use this information to identify patterns and make more accurate and informed economic decisions.

Keywords: Big Data, Econometrics, Information Technology, Machine learning, Forecasting, Financial market analysis, Business analytics.

В современной эпохе цифровой революции объем данных, генерируемых и накапливаемых в различных сферах деятельности, растет в геометрической прогрессии. Эта экспоненциальная величина получила название "большие данные" или "Big Data". Большие данные отличаются от объемов данных, с которыми привыкли работать, и представляют собой объемные, разнообразные, быстро меняющиеся и структурированные или неструктурированные данные, которые требуют новых подходов к их анализу и использованию.

Эконометрика, научная дисциплина, изучающая экономические явления и разрабатывающая методы представления, измерения и анализа

этих явлений, тоже оказывается перед вызовами, представленными большими данными. Стандартные методы анализа, разработанные для работы с меньшими объемами данных, могут быть неэффективны или неприменимы в условиях больших данных. Поэтому разработка информационных технологий, методов анализа и использования больших данных в рамках эконометрических исследований становится неотъемлемой частью развития эконометрики как науки.

Цель данной научной статьи заключается в изучении и обсуждении применения больших данных в контексте эконометрических исследований. Будут рассмотрены различные методы работы с большими данными, их преимущества и ограничения, а также сферы исследования экономики, в которых применение больших данных демонстрирует значимые результаты и влияние. Одновременно обсудим проблемы и ограничения, с которыми исследователи сталкиваются при работе с большими данными, и предложим возможные пути преодоления этих ограничений.

Понимание и использование больших данных в эконометрических исследованиях открывает новые горизонты для построения более точных моделей и прогнозов, выявления скрытых закономерностей, определения причинно-следственных связей и разработки эффективных стратегий принятия решений в сфере экономики. Статья призвана подчеркнуть важность и потенциал использования больших данных в эконометрических исследованиях, а также предложить рекомендации для исследователей и практиков, которые стремятся эффективно использовать эти данные в своей работе.

Для того чтобы более глубоко понять применение больших данных в эконометрике, необходимо разобраться в их определении и особенностях. Определение "большие данные" включает в себя несколько ключевых характеристик, которые являются отличительными чертами данных, требующими специальных подходов и методов анализа [1].

Первым аспектом больших данных является их объем. В то время как в прошлом мы работали с относительно небольшими объемами данных, в настоящее время нам предстоит иметь дело с огромными объемами данных, измеряемыми в терабайтах и петабайтах. Например, социальные сети, финансовые рынки, медицинская диагностика и другие сферы генерируют огромные объемы данных, которые требуют адаптации привычных методов анализа [2].

Второй аспект - разнообразие данных. Большие данные могут быть представлены в различных форматах и структурах. Это могут быть структурированные данные, такие как базы данных и таблицы, а также неструктурированные данные, такие как текстовые документы, аудио- и видеозаписи, изображения и другие форматы. Это включает данные из разных источников - социальных сетей, датчиков, онлайн-платформ и многих других. Разнообразие данных требует различных методов и инструментов для их обработки и анализа [2].

Третья особенность больших данных – это скорость их генерации. Большие данные могут поступать с высокой скоростью и требуют оперативного анализа. Например, финансовые рынки или мониторинг социальных медиа генерируют огромные объемы данных в реальном времени. Для их анализа необходимо обладать методами, которые позволяют обрабатывать и анализировать данные с высокой скоростью [2].

Четвертая особенность больших данных – это достоверность данных. Большие данные могут содержать высокий уровень шума, ошибок и неоднородности в силу разнообразия источников данных. Это может быть вызвано ошибками в сборе данных, неточностью, а также наличием псевдонимов и несоответствующей информации. Поэтому чистка, предварительная обработка и проверка данных на достоверность и качество являются важными шагами при работе с большими данными [2].

Анализ больших данных представляет особые вызовы, требующие специальных методов и техник. Одни стандартные методы и инструменты могут не справиться с работой с большими объемами данных, необходимо разработать новые подходы, обеспечивающие эффективность и точность анализа. В дальнейшей части статьи будут рассмотрены методы работы с большими данными в эконометрике, а также их возможности и ограничения.

Методы машинного обучения предоставляют мощный инструментарий для работы с большими данными в эконометрике. Алгоритмы машинного обучения, такие как регрессия, случайные леса, градиентный бустинг и нейронные сети, могут быть успешно применены для прогнозирования экономических показателей, анализа зависимостей и классификации данных [3]. Они основаны на использовании большого количества наблюдений и могут автоматически обнаруживать скрытые закономерности и сложные взаимосвязи в данных. Однако, для эффективного применения методов машинного обучения необходимо иметь достаточно большой объем данных и высококачественные переменные для обучения моделей.

Методы панельных данных предоставляют возможность анализировать данные, собранные на протяжении времени у нескольких наблюдаемых единиц. Это позволяет учитывать временные и межъявленческие зависимости, а также контролировать эффекты индивидуальных характеристик. Анализ панельных данных может быть осуществлен с применением различных моделей, таких как модели случайных эффектов или фиксированных эффектов. Для работы с большими данными в этом случае может потребоваться использование эффективных вычислительных методов и алгоритмов, таких как дополнительные оценивания и бутстрэп [3].

Методы сжатия данных представляют подходы, направленные на уменьшение размера данных без потери информации. Они основаны на использовании различных методов сжатия, таких как матричная факторизация, техники обрезки и агрегирование данных [3]. Подходы к сжатию данных могут быть полезны для обработки и хранения больших наборов данных, особенно при ограниченных ресурсах вычислительной мощности и памяти. Однако, важно учитывать, что сжатие данных может привести к потере некоторой информации и снижению точности анализа.

Применение этих методов работы с большими данными в эконометрическом анализе демонстрируется во множестве исследований. Например, анализируя огромный объем финансовых данных, исследователи могут использовать методы машинного обучения, чтобы предсказывать цены акций или классифицировать рыночные тренды [3]. Также эконометристы могут применять методы панельных данных для анализа влияния экономических политик на рост внутреннего потребления в различных странах. Методы сжатия данных могут быть полезны при анализе больших временных рядов, таких как данные о климатических изменениях.

Эконометрика использует статистические методы и математические модели для изучения экономических явлений и анализа данных [4]. Введение больших данных в область эконометрики привело к расширению возможностей и повышению точности анализа.

Одной из областей, где применение больших данных оказывает значимое влияние, является анализ финансовых рынков. Благодаря большому объему доступных данных, исследователи могут более точно анализировать колебания на рынках, выявлять тенденции и прогнозировать изменения ценовых индексов и активов. Такие данные могут включать в себя информацию о торговле на биржах, финансовых новостях и других факторах, которые могут влиять на финансовые рынки.

Прогнозирование экономических показателей также получило значительное преимущество от применения больших данных. Благодаря анализу обширных данных, экономисты и финансовые аналитики могут строить более точные прогнозы по таким важным переменным, как ВВП, инфляция, безработица и другие экономические показатели [4]. Это помогает руководителям принимать обоснованные решения и позволяет лучше планировать экономическую деятельность.

Макроэкономический анализ также сильно выиграл от использования больших данных. Изучение экономических данных в масштабе стран и регионов позволяет выявить взаимодействия и зависимости между различными экономическими факторами. Аналитики могут использовать данные о ВВП, инвестициях, торговле, безработице и других показателях для определения тенденций и стратегического планирования [4].

Исследование рыночной конкуренции также может получить значительные преимущества от анализа больших данных. Опираясь на информацию о ценах, спросе, продажах и других факторах рынка, исследователи могут выявить тренды конкуренции и определить факторы, влияющие на успех предприятий на рынке [4]. Это позволяет разрабатывать эффективные стратегии маркетинга и ценообразования.

Однако, при использовании больших данных в эконометрике возникают и некоторые вызовы. Работа с большим объемом данных требует достаточно мощных вычислительных ресурсов и экспертизы в обработке данных. Также важно учитывать приватность данных и обеспечивать их защиту при сборе и анализе.

Анализ больших данных в эконометрике представляет ряд проблем и ограничений, которые необходимо учитывать при проведении исследований.

Один из основных нюансов состоит в обработке и хранении больших объемов данных. Большие данные требуют мощных вычислительных ресурсов и эффективных алгоритмов для их анализа. Для преодоления этого можно использовать параллельные вычисления и распределенные системы хранения данных [5]. Градиентный бустинг строится последовательно, и каждая базовая модель обучается на ошибках предыдущей. Однако, можно ускорить обучение с использованием параллельных вычислений, обучая некоторые базовые модели одновременно.

Предположим, у нас есть N базовых моделей. Тогда прогноз градиентного бустинга можно переписать следующим образом:

$$F(X)=\sum_{m=1}^{M}\tau_m h_m(X)=\sum_{i=1}^{N}\sum_{m=1}^{M/N}\tau_{i,m}h_{i,m}(X),$$ где:

$\tau_{i,m}$ – коэффициент обучения для *i*-ой базовой модели и *m*-ой итерации,

$h_{i,m}(X)$ – прогноз *i*-ой базовой модели на *m*-ой итерации [5].

Также важно разработать оптимальные методы для фильтрации и сжатия данных, чтобы уменьшить их объем, не потеряв при этом важные статистические свойства.

Еще одним вызовом является проблема выборки и представительности. Большие данные могут быть нерепрезентативными из-за их объема и разнообразия источников. Важно разработать методы, которые учитывают эту проблему и позволяют проводить анализ на основе случайной выборки, чтобы получить статистически значимые результаты. Также нужно аккуратно оценивать степень случайности и систематической ошибки в данных, чтобы избежать неправильной интерпретации результатов.

Проблемы причинности и интерпретации результатов также являются важными задачами в анализе больших данных. Большой объем данных

может позволить нам найти статистически значимые связи между переменными, но это не означает, что они являются причинно-следственными. Для преодоления этой проблемы необходимо применять методы эконометрики, которые учитывают эндогенность переменных и проверяют гипотезы о причинно-следственной связи [5]. Также важно предоставлять интерпретируемые результаты, которые можно объяснить экономической теорией и практическими выводами. В качестве примера предположим, у нас есть модель простой линейной регрессии:

$$Y=\beta_0+\beta_1 X+\varepsilon, \text{ где:}$$

Y – зависимая переменная,

X – эндогенная независимая переменная,

β_0 – коэффициент сдвига,

β_1 – коэффициент наклона,

ε - случайная ошибка.

Однако, если X эндогенна (коррелирует с ошибкой), оценки коэффициентов могут быть смещены. Для устранения эндогенности, мы используем инструментальную переменную Z, которая коррелирует с X, но не коррелирует с ошибкой.

$$X=\pi_0+\pi_1 Z+u, \text{ где:}$$

Z – инструментальная переменная,

π_0 – коэффициент сдвига для инструмента,

π_1 – коэффициент наклона для инструмента,

u – ошибка инструмента.

Теперь мы можем подставить X из второго уравнения в первое:

$$Y=\beta_0+\beta_1(\pi_0+\pi_1 Z+u)+\varepsilon$$

$$Y=(\beta¿¿0+\beta_1\pi_0)+(\beta¿¿1\pi_1)Z+(\beta¿¿1u+\varepsilon)¿¿¿$$

Теперь, если инструмент Z удовлетворяет условиям инструментальной переменной (корреляция с X и некоррелирован с ошибкой), то мы можем использовать его для получения состоятельных и эффективных оценок коэффициентов β_0 и β_1.

Это демонстрирует простой пример применения инструментальных переменных для учета эндогенности в эконометрике. В более сложных моделях используются различные статистические тесты и дополнительные инструменты для обеспечения справедливости условий использования инструментальных переменных.

Для преодоления этих вызовов является важным развитие новых методов и подходов в эконометрике. Использование машинного обучения и искусственного интеллекта может помочь анализировать большие объемы данных и находить скрытые закономерности. Однако, необходимо учитывать, что эти методы имеют свои ограничения и требуют аккуратного подхода при их применении.

В течение последних десятилетий информационные технологии претерпели значительные изменения, что привело к появлению большого объема данных, которые могут быть использованы для эконометрических исследований. Однако, для того чтобы эти данные были полезными, требуется развитие соответствующих методов анализа.

В целом, развитие информационных технологий и методов анализа является важным фактором для успешного использования больших данных в эконометрических исследованиях. Понимание потенциала и ограничений больших данных поможет исследователям и практикам принять взвешенные решения и сделать значимый вклад в экономическую науку и практику.

Использованные источники:

1. Mayer-Schönberger V., Cukier K. – Big Data: A Revolution That Will Transform How We Live, Work, and Think/ EAMON DOLAN BOOK, 2013. – 272 c.
2. Hastie T., Tibshirani R., Friedman J. – The Elements of Statistical Learning: Data Mining, Inference, and Prediction/ Springer, 2016 – 767 c.
3. Liebowitz J. – Big Data and Business Analytics/ Auerbach Publications, 2017 – 304 c.
4. Newbold P., Carlson W., Thorne B. – Econometrics: Statistical Foundations and Applications/ Springer, 2014 – 592 c.
5. Marr B. – Big Data: Using Smart Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance/ Wiley, 2013 – 256 c.